User Association for Backhaul Load Balancing with Quality of Service Provisioning for Heterogeneous Networks

Ying Loong Lee, Teong Chee Chuah, Ayman A. El-Saleh, and Jonathan Loo

***Abstract*—This paper addresses the user association problem for quality of service (QoS) provisioning and backhaul load balancing (LB) in heterogeneous networks (HetNets). This problem is exacerbated by base stations with different backhaul capacities and users with diverse QoS requirements. A user association scheme is proposed to achieve QoS provisioning and backhaul LB for HetNets. Simulation results show that the proposed scheme outperforms conventional user association schemes in terms of call blocking probability, QoS, and backhaul LB.**

***Index Terms*—Load balancing, quality of service, user association, heterogeneous networks.**

I. Introduction

HETEROGENEOUS networks (HetNets) have emerged as a promising paradigm to boost user capacity and data rates, thus regarded as one of the key architectures of fifth generation (5G) systems. In HetNets, small-cell base stations (BSs) are deployed within macrocells to improve the coverage of the areas which are poorly served. However, challenges in terms of interference, fairness and quality of service (QoS) arise in HetNet deployment [1]. Many studies have been done to tackle the challenges.

HetNets have received significant attention due to its advantages towards realizing smart cities and internet of things (IoT) networks. In particular, multi-operator network sharing and slicing has recently been considered as the key feature of HetNets for supporting smart city and IoT applications. Many researchers have delved into this research topic from the resource allocation perspective, and proposed new resource allocation architectures for HetNets to implement multi-operator network sharing with the objective to support multimedia applications [2]–[5]. Meanwhile, several researchers addressed the technical challenges related to HetNets from the load balancing (LB) perspective [6]–[10]. This research direction has become increasingly important because user association, which is the key LB mechanism, is usually performed before resource allocation and thus it can greatly affect the performance of the latter. Thus, improper user association not only leads to load imbalance but inefficient resource allocation among users, hence limiting QoS provisioning which is crucial for IoT-based multimedia applications. Despite the existing studies on load balancing for HetNets, some open issues remain unresolved, thus motivating the current study.

In existing studies, the number of users served [6], or the amount of resources consumed [6], [7] are often considered as the load carried by each BS. However, the backhaul capacity, which is in fact the bottleneck of the load that can be carried by each BS, is not considered. Hence, the conventional LB techniques in [6]–[9] cannot be directly applied for backhaul LB. The backhaul LB problem becomes more challenging when each BS in the HetNets has a different backhaul capacity. Although [10] has addressed backhaul LB, the diverse QoS requirements of users have not been considered. For instance, users with high data rate requirements may experience starvation if they are offloaded to a BS with low backhaul capacity.

The current study aims to investigate backhaul LB by taking into account the diverse backhaul capacity of each small-cell and users' QoS requirements. This study focuses on downlink backhaul LB and is based on 3GPP Long Term Evolution (LTE). The contributions of this study are summarized as follows: 1) An optimization problem is formulated to maximize a logarithmic utility function with respect to the backhaul load efficiency of small-cells in order to achieve proportional fairness, subject to the QoS requirements of each user and the backhaul capacity of each small-cell; 2) two algorithms are derived based on dual decomposition with one implemented at the user side and another implemented at the BS side; 3) the performance of the proposed scheme is compared with several conventional LB schemes in terms of QoS provisioning and fairness.

II. System Model and Problem Formulation

An LTE-based HetNet consisting of a macrocell BS (MBS) and several small-cell BSs as shown in Fig. 1 is considered. $\mathcal{S}$ and $\mathcal{U}$ denote the sets of BSs (with $s = 0$ denoting the MBS) and user equipment (UEs), respectively. The number of available physical resource blocks (PRBs) in the HetNet is denoted by K and full PRB reuse is allowed in the HetNet. b_{su} is defined as the association indicator where $b_{su} = 1$ if UE u associates with BS s, else $b_{su} = 0$. The data rate of UE u achieved on a PRB by BS s is modeled as Shannon's capacity:

This work was supported in part by Telekom Malaysia Research and Development under grant RDTC/160915.

Y. L. Lee was with Xiamen University Malaysia, 43900 Sepang, Selangor, Malaysia. He is now with the Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Cheras 43000, Kajang, Selangor, Malaysia (e-mail: yingloonglee@gmail.com).

T. C. Chuah is with the Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia (e-mail: tcchuah@mmu.edu.my).

A. A. El-Saleh is with the College of Engineering, A'Sharqiyah University, Ibra 400, Oman (e-mail: ayman.elsaleh@asu.edu.om).

J. Loo is with the School of Computing and Engineering, University of West London, London W5 5RF, United Kingdom (e-mail: jonathan.loo@uwl.ac.uk).

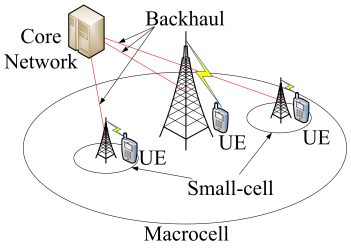


Fig. 1. System model of HetNet.

$$R_{su} = B\log(1+\Gamma_{su}) \tag{1}$$

where B is the bandwidth of a PRB, and

$$\Gamma_{su} = \frac{P_s G_{su}}{\sum_{i\in\mathcal{S}\setminus\{s\}} P_i G_{iu} + P_{\text{AWGN}}} \tag{2}$$

is the signal-to-interference-plus-noise-ratio (SINR) between BS s and UE u. In (2), P_s is the transmission power of BS s, G_{su} is the downlink channel gain between BS s and UE u, P_{AWGN} is the additive white Gaussian noise power. Since user association is assumed to be carried out in a larger time scale, G_{su} is assumed to have been averaged within the association period and over all PRBs in the whole channel bandwidth, i.e., fast fading and frequency-selective fading are averaged out. Therefore, G_{su} is constant regardless of the dynamic channel variations within the association period and the SINR between BS s and UE u for each PRB is the same. Similar SINR models have been adopted in [6]–[8], [11]. Each UE u needs to achieve a target data rate $R_{\text{req},u}$ to meet its QoS requirement. The number of PRBs required by each UE u to meet its $R_{\text{req},u}$ if it is associated with BS s can be estimated as [7]

$$N_{su} = \left\lceil \frac{R_{\text{req},u}}{R_{su}} \right\rceil \tag{3}$$

where $\lceil . \rceil$ denotes the ceiling operator. The backhaul load can be mathematically represented as the ratio of the data rate carried by the backhaul link to the backhaul capacity:

$$\eta_s = \frac{\sum_{u\in\mathcal{U}} b_{su} N_{su} R_{su}}{C_{\text{bh},s}} \tag{4}$$

where $C_{\text{bh},s}$ is the backhaul capacity of BS s. Further, the total number of PRBs required by BS s to serve its associated UEs can be determined as

$$M_s = \sum_{u\in\mathcal{U}} b_{su} N_{su}. \tag{5}$$

The backhaul LB problem can be formulated as a network utility maximization problem whereby maximizing the utility function would lead to fairness. A suitable utility function is the logarithmic utility function with respect to η_s which leads to diminishing returns and thus encourages LB [6]. However, a difficulty arises, where some UEs may associate with BSs that provide low channel quality due to path loss and fading. Therefore, it is imperative to associate UEs with BSs that provide high channel quality. Thus, the logarithmic utility function with respect to the backhaul load efficiency is maximized, where the backhaul load efficiency of BS s is defined as

$$\gamma_s = \frac{\eta_s}{M_s}. \tag{6}$$

By maximizing the logarithmic utility function with respect to γ_s, backhaul LB can be achieved while encouraging UEs to associate with BSs that provide high channel quality. The user association problem can be formulated as follows:

$$\max_{\mathbf{b}} \sum_{s\in\mathcal{S}} \log \gamma_s \tag{7}$$

subject to

$$\eta_s \le 1 \quad \forall s \in \mathcal{S} \tag{7a}$$
$$M_s \le K \quad \forall s \in \mathcal{S} \tag{7b}$$
$$\sum_{s\in\mathcal{S}} b_{su} = 1 \quad \forall u \in \mathcal{U} \tag{7c}$$
$$b_{su} \in \{0,1\} \quad \forall s \in \mathcal{S}, u \in \mathcal{U} \tag{7d}$$

Constraint (7a) ensures that the total data rate achieved by each BS does not exceed its backhaul capacity. Constraint (7b) guarantees that the total number of PRBs allocated by each BS to all its associated UEs does not exceed the maximum number of available PRBs. Constraint (7c) ensures that each UE can only associate with one BS.

III. Proposed Backhaul Load Balancing Scheme

The problem in (7) can be classified as a 0-1 integer programming problem, which is generally difficult to solve. Methods such as the branch-and-bound approach would take exponential time complexity in the worst case to obtain the optimal solution, which is impractical for modest or large networks. To make (7) more tractable, constraint (7d) is relaxed to continuous values: $0 \le b_{su} \le 1$, which makes (7) convex. Also, the objective function in (7) can be rewritten as

$$\sum_{s\in\mathcal{S}} \log \gamma_s = \sum_{s\in\mathcal{S}} \log\left(\frac{\eta_s}{M_s}\right) = \sum_{s\in\mathcal{S}} \log \eta_s - \sum_{s\in\mathcal{S}} \log M_s.$$

Thus, with relaxation of (7d), (7) can be re-expressed as

$$\max_{\mathbf{b}} \left(\sum_{s\in\mathcal{S}} \log \eta_s - \sum_{s\in\mathcal{S}} \log M_s \right) \tag{8}$$

subject to (7a)-(7c) and

$$0 \le b_{su} \le 1 \quad \forall s \in \mathcal{S}, u \in \mathcal{U}. \tag{8a}$$

To solve (8), two sets of new variables, i.e., $\mathbf{x}$ and $\mathbf{y}$ where $x_s = \eta_s$ and $y_s = M_s$. This allows (8) to be transformed into

$$\max_{\mathbf{b},\mathbf{x},\mathbf{y}} \left(\sum_{s\in\mathcal{S}} \log x_s - \sum_{s\in\mathcal{S}} \log y_s \right) \tag{9}$$

subject to (7c), (8a), and

$$x_s = \eta_s \quad \forall s \in \mathcal{S} \tag{9a}$$
$$y_s = M_s \quad \forall s \in \mathcal{S} \tag{9b}$$
$$0 < x_s \le 1 \quad \forall s \in \mathcal{S} \tag{9c}$$
$$0 < y_s \le K \quad \forall s \in \mathcal{S}. \tag{9d}$$

Next, the problem in (9) can be solved by dual decomposition. Firstly, the partial Lagrangian of (9) can be written as

$$\begin{aligned}\mathcal{L}(\mathbf{b},\mathbf{x},\mathbf{y},\boldsymbol{\alpha},\boldsymbol{\beta}) = &\sum_{s\in\mathcal{S}} \log x_s - \sum_{s\in\mathcal{S}} \log y_s \\ &+ \sum_{s\in\mathcal{S}} \alpha_s(\eta_s - x_s) + \sum_{s\in\mathcal{S}} \beta_s(y_s - M_s)\end{aligned} \tag{10}$$

where α_s and β_s are the Lagrange multipliers corresponding to constraints (9a) and (9b), respectively. The corresponding dual function can be expressed as

$$\mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{cases} \max_{\mathbf{b},\mathbf{x},\mathbf{y}} \mathcal{L}(\mathbf{b}, \mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \text{subject to (7c), (8a), (9c) and (9d).} \end{cases} \tag{11}$$

In fact, the dual function in (11) can actually be written as

$$\mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathcal{D}_1(\boldsymbol{\alpha}) + \mathcal{D}_2(\boldsymbol{\beta}) + \mathcal{D}_3(\boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{12}$$

where

$$\mathcal{D}_1(\boldsymbol{\alpha}) = \max_{0<\mathbf{x}\leq 1} \sum_{s\in\mathcal{S}} (\log x_s - \alpha_s x_s), \tag{13}$$

$$\mathcal{D}_2(\boldsymbol{\beta}) = \max_{0<\mathbf{y}\leq K} \sum_{s\in\mathcal{S}} (\beta_s y_s - \log y_s), \tag{14}$$

$$\mathcal{D}_3(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{cases} \max_{\mathbf{b}} \sum_{s\in\mathcal{S}} (\alpha_s \eta_s - \beta_s M_s) \\ \text{subject to (7c) and (8a).} \end{cases} \tag{15}$$

The dual problem of (9) can be expressed as

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathcal{D}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{16}$$

where its solution is also the one to (9).

As the problems in (13) and (14) are convex, the solutions can be obtained by setting their derivatives to zero:

$$x_s = \min\left(1, \frac{1}{\alpha_s}\right), y_s = \min\left(K, \frac{1}{\beta_s}\right). \tag{17}$$

The problem in (15) can be solved using the Karush-Kuhn-Tucker (KKT) conditions [12], which can be obtained by differentiating the partial Lagrangian of (15) with respect to b_{su}. Then, the following solution can be obtained analytically from the KKT conditions:

$$b_{su} = \begin{cases} 1 & s = \arg\max_{i\in\mathcal{S}} N_{iu} \left(\frac{\alpha_i R_{iu}}{C_{\text{bh},i}} - \beta_i\right) \\ 0 & \text{Otherwise} \end{cases} \forall u \in \mathcal{U}. \tag{18}$$

It is noteworthy that (18) gives a binary solution of b_{su}, which satisfies constraints (7c) and (7d), and thus no additional step is needed to restore b_{su} to a binary value. Then, the dual problem in (16) can be solved using the subgradient method [13]. Since $x_s > 0$ and $y_s > 0$, α_s and β_s must be nonnegative for the solutions in (17), thus the following projected subgradient method is used to update α_s and β_s such that their values fall within the range of nonnegative values:

$$\alpha_s^{(t+1)} = \left[\alpha_s^{(t)} - \delta(\eta_s - x_s)\right]^+, \tag{19}$$

$$\beta_s^{(t+1)} = \left[\beta_s^{(t)} - \delta(y_s - M_s)\right]^+, \tag{20}$$

where $[z]^+ = \max(0, z)$, δ is the square summable but nonsummable step size, and t is the iteration index. After the subgradient updates, the process is repeated until convergence or it reaches the maximum number of iterations $T_{\max}$.

The proposed solution can be implemented in a distributed manner among UEs and BSs. The proposed scheme consists of Algorithm 1, which is implemented at the UE side, and Algorithm 2 which is implemented at the BS side. These algorithms will be executed at the UE and BS sides until α_s and β_s converge within a very small tolerance ϵ, or $T_{\max}$ has been reached. Similar convergence conditions have been used in [14] and [15]. In each iteration, the complexity of Algorithm 1 and Algorithm 2 is both $O(|\mathcal{S}||\mathcal{U}|)$, because $|\mathcal{S}||\mathcal{U}|$ calculations are needed to update $\mathbf{b}$ at the UE side and $|\mathcal{S}||\mathcal{U}|$ calculations are needed to update $\mathbf{b}$, $\mathbf{x}$ and $\mathbf{y}$ at the BS side. Thus, the total complexity of Algorithms 1 and 2 for the entire process is both $O(T_{\max}|\mathcal{S}||\mathcal{U}|)$. It is noteworthy that the solution obtained from Algorithms 1 and 2 is optimal to (9) but it may not be optimal or even feasible to the original user association problem in (7) due to the relaxation of b_{su}. To ensure that the solution is feasible, especially the fulfillment of constraints (7a) and (7b), a UE dropping mechanism is introduced at the BS, whereby UEs with excessive resource demands will first be dropped. If constraints (7a) and (7b) are not satisfied at the BS, one or more random UEs will be dropped until (7a) and (7b) are satisfied.

Algorithm 1 Operation at UE side in each iteration

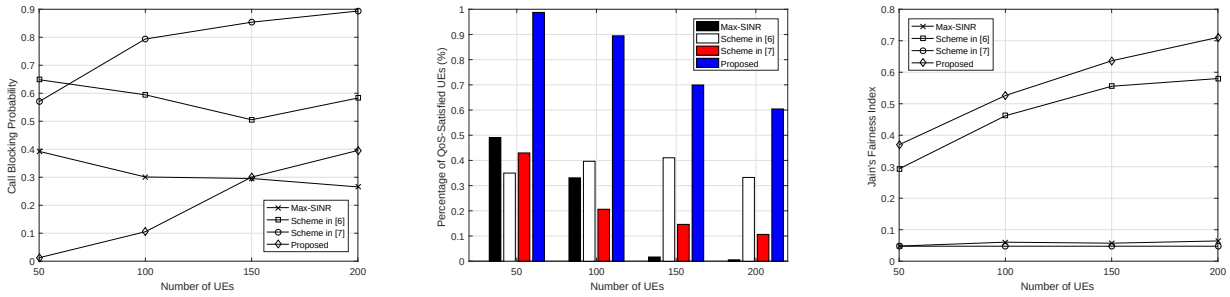
1: Initialize $t = 0$; each UE u measures the SINR based on the pilot signal from each BS s, and estimates R_{su} and N_{su} with (1) and (3), respectively.
2: Each UE u sends the information of R_{su} and N_{su} to each BS s.
3: **repeat**
4: Each UE u receives the values of α_s, β_s and $C_{\text{bh},s}$ from each BS s via BS broadcast.
5: Each UE u determines the target BS s to be associated with according to (18).
6: Each UE u sends the user association request to the chosen target BS s.
7: $t \leftarrow t + 1$.
8: **until** UE u receives association confirmation from the target BS.

Algorithm 2 Operation at BS side in each iteration

1: Initialize $t = 0$; each BS s initializes α_s and β_s and broadcast α_s, β_s and $C_{\text{bh},s}$ to the network.
2: Each BS s receives the values of R_{su} and N_{su} from each UE u.
3: **repeat**
4: Each BS s receives the user association requests from UEs and updates the corresponding b_{su} from the request information.
5: Each BS s updates x_s and y_s using (17).
6: Each BS s updates α_s and β_s with (19) and (20) respectively.
7: Each BS s broadcast the updated α_s and β_s, as well as $C_{\text{bh},s}$.
8: $t \leftarrow t + 1$.
9: **until** $\left|\alpha_s^{(t+1)} - \alpha_s^{(t)}\right| < \epsilon$ and $\left|\beta_s^{(t+1)} - \beta_s^{(t)}\right| < \epsilon$, or $t = T_{\max}$.
10: Each BS s sends association confirmation to the requested UEs.

IV. Numerical Results and Discussion

A macrocell of 1 km radius that is overlaid with 20 randomly located small-cells is considered. The transmission power of the MBS and SBSs are set to 43 dBm and 20 dBm respectively. The backhaul capacity of the MBS is set to 50 Mb/s whereas that of the SBSs are randomly set within [1, 5] Mb/s. UEs are randomly distributed within the macrocell and their required data rate are randomly set within [300, 500] kb/s. The channel consists of 100 PRBs with each having 180 kHz bandwidth. The following path loss models: $128.1 + 37.6\log(d)$ (dB) and $127 + 30\log(d)$ (dB) are used for the macrocell and small-cell, respectively, where d (km) is the distance between the UE and the BS. A channel with zero-mean unit-variance Rayleigh fading and zero-mean log-normal shadowing with 10-dB standard deviation is considered. The noise figure and noise spectral density are set to 9 dB and -174



(a) (b) (c)

Fig. 2. (a) Call blocking probability; (b) percentage of QoS-satisfied UEs; (c) backhaul load balancing performance.

dBm/Hz, respectively. Also, the proposed scheme is compared with the maximum SINR (max-SINR)-based user association scheme with equal resource allocation, and the schemes in [6] and [7]. The simulation results are averaged over 100 instances with each having UEs being stationary but located at different positions and experiencing different channel conditions.

The convergence behavior of the proposed scheme of up to 1000 iterations has been analyzed and it is observed that the proposed scheme converges within 100 iterations at 50 UEs and within 1000 iterations at 100 UEs. However, it does not converge within 1000 iterations at 150 and 200 UEs due to the increasing numbers of UEs which increases the problem size. Nonetheless, the proposed scheme can still achieve substantial performance gains compared to the existing schemes even though the proposed scheme has not achieved convergence. For the subsequent results, $T_{\max}$ is set to 100.

In Fig. 2(a), the call blocking probability [7] defined as the ratio of the number of dropped UEs to the total number of UEs is evaluated. The proposed scheme achieves lower call blocking probabilities than the schemes in [6] and [7], because the proposed scheme has taken into account both the backhaul capacity as well as QoS requirements of the UEs, unlike the other two schemes. Notably, the max-SINR scheme achieves lower blocking probabilities at 150 and 200 UEs than other three schemes because it accepts UEs without considering whether their QoS requirements can be satisfied.

Fig. 2(b) shows the percentage of QoS-satisfied UEs associated with the BSs in the HetNet. The proposed scheme is shown to allow more UEs to achieve their data rates compared with the other three schemes, because the max-SINR scheme does not take into account QoS requirements of the UEs whereas the schemes in [6] and [7] have more UEs dropped as shown in Fig. 2(a), resulting in fewer QoS-satisfied UEs.

In Fig. 2(c), Jain's fairness index [16] defined as $\frac{\left(\sum_{s\in\mathcal{S}}\eta_s\right)^2}{|\mathcal{S}|\sum_{s\in\mathcal{S}}\eta_s^2}$ is used to evaluate the backhaul LB performance of the HetNet. The proposed scheme is shown to outperform the other schemes because the former takes into account the limited backhaul capacity in the LB process.

V. Conclusion

A user association scheme for backhaul LB with QoS provisioning in HetNets is presented. Simulation results show that the proposed scheme outperforms the existing user association schemes in terms of call blocking probability, QoS and fairness.

References

[1] Y. L. Lee, T. C. Chuah, J. Loo, and A. Vinel, "Recent advances in radio resource management for heterogeneous LTE/LTE-A networks," *IEEE Commun. Surveys. Tuts.*, vol. 16, no. 4, pp. 2142–2180, Jun. 2014.

[2] R. Kunst, L. Avila, E. Pignaton, S. Bampi, and J. Rochol, "Improving QoS in multi-operator cellular networks," in *Proc. IEEE 12th WiMob*, New York, USA, Oct. 2016, pp. 1–8.

[3] ——, "A resources sharing architecture for heterogeneous wireless cellular networks," in *Proc. IEEE 41st LCN*, Dubai, United Arab Emirates, Nov. 2016, pp. 228–231.

[4] ——, "Improving network resources allocation in smart cities video surveillance," *Comput. Netw.*, vol. 134, pp. 228–244, Apr. 2018.

[5] Y. L. Lee, J. Loo, T. C. Chuah, and L.-C. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2146–2161, Apr. 2018.

[6] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[7] T. Zhou, Y. Huang, W. Huang, S. Li, Y. Sun, and L. Yang, "QoS-aware user association for load balancing in heterogeneous cellular networks," in *Proc. IEEE 80th VTC-Fall*, Vancouver, BC, Canada, Sep. 2014, pp. 1–5.

[8] E. Rakotomanana and F. Gagnon, "Fair load balancing in heterogeneous cellular networks," in *Proc. IEEE ICUWB*, Montreal, QC, Canada, Oct. 2015, pp. 1–5.

[9] I. Sohn and S. H. Lee, "Distributed load balancing via message passing for heterogeneous cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 9287–9298, Nov. 2016.

[10] Y. Xu, R. Yin, and G. Yu, "Adaptive biasing scheme for load balancing in backhaul constrained small cell networks," *IET Commun.*, vol. 9, no. 7, pp. 999–1005, Apr. 2015.

[11] N. Trabelsi, C. C. Chen, R. E. Azouzi, L. Roullet, and E. Altman, "User association and resource allocation optimization in LTE cellular networks," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 2, pp. 429–440, Jun. 2017.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge Univ. Press, 2004.

[13] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," Stanford University, 2006-07, notes for EE364b. [Online]. Available: https://web.stanford.edu/class/ee392o/subgrad_method.pdf

[14] H. Zhang, Y. Liu, and M. Tao, "Resource allocation with subcarrier pairing in OFDMA two-way relay networks," *IEEE Wireless Commun. Lett.*, vol. 1, no. 2, pp. 61–64, Apr. 2012.

[15] T. Wang and L. Vandendorpe, "Iterative resource allocation for maximizing weighted sum min-rate in downlink cellular OFDMA systems," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 223–234, Jan. 2011.

[16] R. Jain, *The Art of Computer Systems: Performance Analysis*. Hoboken, NJ, USA: John Wiley & Sons, 1991.